Restaurant Visitor Forecasting

Team name - NSA

Ankita Paul
MT2020053
ankita.paul@iiitb.org

Neha Kothari
MT202010
neha.kothari@iiitb.org

Soham Chatterjee
MT2020071
soham.chatterjee@iiitb.org

Abstract—**Running a thriving local restaurant isn't always as charming as first impressions appear. There are often all sorts of unexpected troubles popping up that could hurt business.**

In this competition, we are asked to use reservation and visitation data to predict the total number of visitors to a restaurant for future dates. This information is supposed to help restaurants be much more efficient and allow them to focus on creating an enjoyable dining experience for their customers. We are also given some additional metadata on the restaurants such as genre and location.

Index Terms - **Feature Engineering, Linear Regression, K Nearest Neighbours, Light Gradient Boosting Machine, Decision Tree, Random Forest Regressor.**

I. Introduction

One common predicament is that restaurants need to know how many customers to expect each day to effectively purchase ingredients and schedule staff members. This forecast isn't easy to make because many unpredictable factors affect restaurant attendance, like festivities and local competition. It's even harder for newer restaurants with little historical data.

One could argue that the visitors forecasting can be solved using simple statistics or the restaurant owner has a general idea of daily visitor trends and with practice, accurate prediction is possible. In both cases, the experience, and historical data is necessary to arrive at a number. However, this isn’t the case with new restaurants with no historical data. ML accounts different factors and conditions such as local competition and time of the year before spilling a number, even for restaurants with less or no historical data. With ML the confidence of tackling this problem is higher compared to the arguments.

With the help of our visualization and model we aim to help predict the number of visitors for different restaurants taking into account the genre, area, time of the year, etc.

The rest of the paper proceeds as follows: In Sec. 2 we describe our dataset and the evaluation criteria, Sec. 3 covers our visualizations, EDA and their inferences Sec. 4 will discuss about Data pre-processing and feature extraction, Sec. 5 will discuss training methods and comparison of our model with the various other approaches.

II. Dataset and Evaluation Metric

A. Dataset

The data comes from two separate sites:

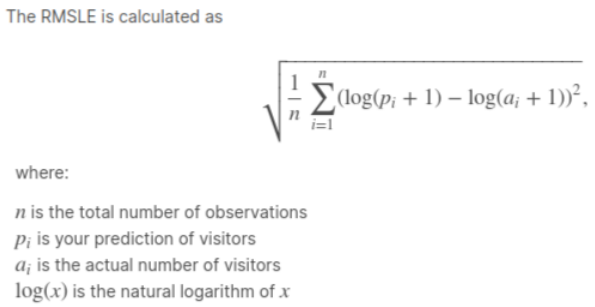
i. Hot Pepper Gourmet (HPG): similar to Yelp, here users can search restaurants and also make a reservation online

ii. AirREGI / Restaurant Board (air): similar to Square, a reservation control and cash register system.

Both provide information about the genre, area, latitude, longitude, reserve date-time and visit date time of the different stores.

The training data covers the dates from 2016 until early (first week) April 2017. The test set covers the mid weeks (second and third weeks) of April 2017. The training and testing set both omit days where the restaurants were closed. Training dataset has information about AIR store IDs, dates and the corresponding actual number of visitors.

B. Evaluation Metric



The RMSLE is calculated as

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(p_i+1)-\log(a_i+1))^2},$$

where:

n is the total number of observations
p_i is your prediction of visitors
a_i is the actual number of visitors
$\log(x)$ is the natural logarithm of x

Why RMSLE?: We are trying to help the restaurants to be well prepared for the visitors. If they are underprepared, which is the case when the prediction is less than the actual number, then they are short on resources and that is a bad dining experience for the customers. On the other hand, if the restaurants are over-prepared, as in they bought resources for 20 customers, and only 15 showed up. They still can store the extra resources for the next day, while none of the customers returned unhappily.

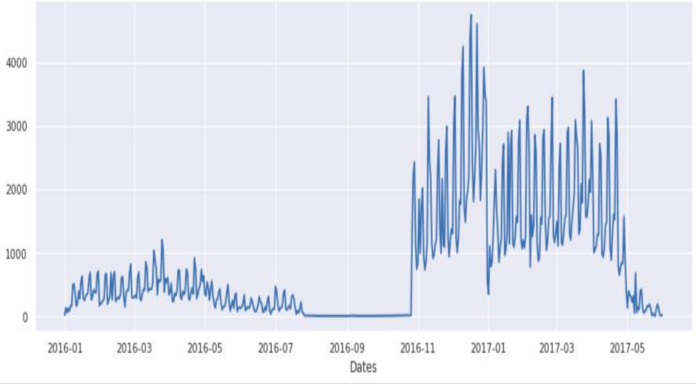
This is exactly why RMSLE is important, which penalizes higher to under predictions compared to over-predictions

III. Visualizations and EDA

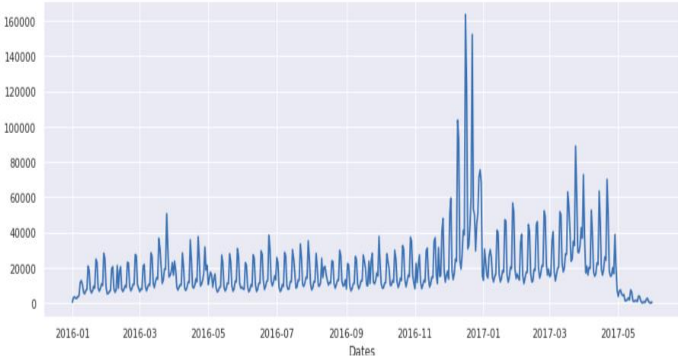
A. Reservations

Reservations made through the 2 different systems

Reservations made through AIR:



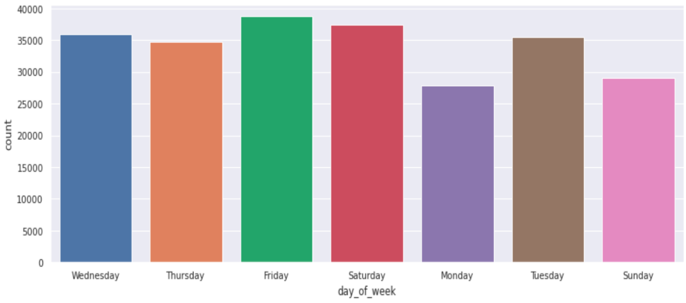
Reservations made through HPG:



There were much fewer reservations made in 2016 through the air system; even none at all for a long stretch of time. The volume only increased during the end of that year. In 2017 the visitor numbers stayed strong. The artificial decline we see after the first quarter is most likely related to these reservations being at the end of the training time frame, which means that long-term reservations would not be part of this data set.

Also during the year there is a certain amount of variation. Dec appears to be the most popular month for restaurant visits. The period of Mar - May is consistently busy.

B. *Day of Week Trends*

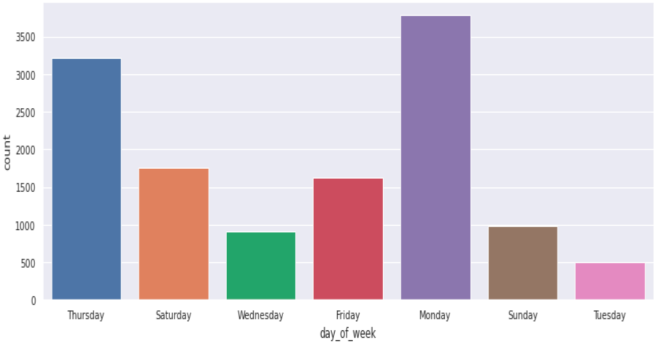


The graph shows the day of the week vs the number of visitors on the day.

Fridays and the weekend appear to be the most popular days; which is expected. Monday has the lowest number of average visitors.

C. *Holiday Trends*

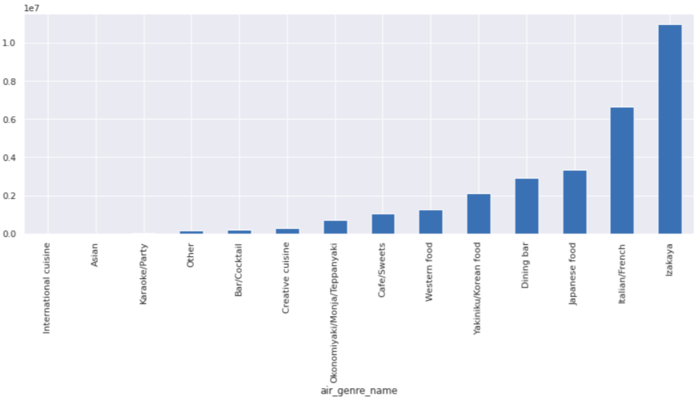
The graph number of visitors and the day of week. The difference is that the number of visitors is only for the holidays.



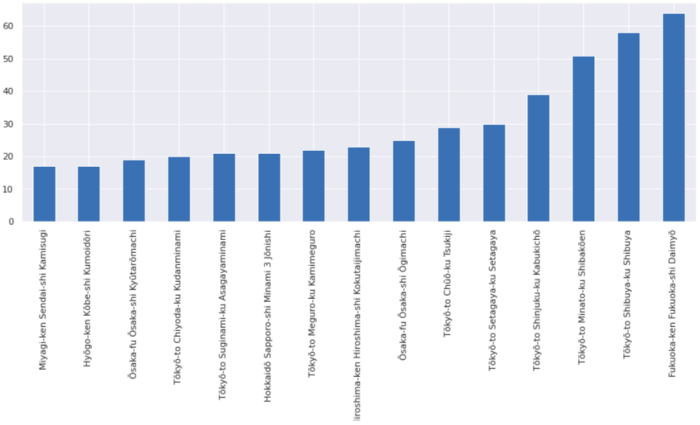
While a weekend holiday has little impact on the visitor numbers, and even decreases them slightly, there is a much more pronounced effect for the weekdays; especially Monday and Tuesday.

D. *Genre and Area*

Visitor Trend for different genres:



Top 15 areas with maximum number of restaurants:



Using the information on the number of genres in each area we can now proceed to quantify the clustering, or “crowdedness”, of our data set and relate it to the visitor numbers. The number of restaurants of a particular genre in an area could help us understand the local competition too.

IV. Feature Engineering

The next step is to derive new features from the existing ones and the purpose of these new features is to provide additional predictive power for our goal to forecast visitor numbers. Those new features can be as simple as deriving the day of the week or the month from a date column.

We converted the “visit_date” column into a python date-time object and extracted key features like month, day-of-week, day of the month and week of the year. These columns are really helpful in getting intrinsic data like patterns in daily visitor numbers, patterns in weekend/weekday visitor numbers, as seen in the data visualization section.

Further, there could be restaurants that are popular for some of their seasonal dishes or weekly menus, so we decided to group the data by (store_id, month) and (store_id, day of week) and derive the mean, median and maximum number of visitors.

We derived the hour gap between the time of reservation and visit time to understand any underlying pattern and categorized them into different buckets.

It’s usually common for employees to relax if the next day is a holiday, and maybe dine out. So, we added a feature - ‘holiday_eve’ flag based on the ‘holiday_flag’ column in the calendar info data.

Another feature based on the holiday flag was ‘non_working_day’, where the flag was true for all public holidays and weekends. Some restaurants may have more visitors on a working day and some may expect more on non-working days, so the data was grouped by (store_id, non_working_day) and the mean, median and maximum number of visitors was deduced.

The total number of restaurants in an area and the count of the same genre restaurants in a particular area could help define the local competition and aid in prediction.

For area competition - group the data by area names to get the count of restaurants in the area of the restaurant.

Since each genre has its own trend of visitors, Group the data by area name, genre name to get the count of same genre restaurants in the same area of the restaurant to get the ‘genre competition.’

Area name could be very important information in determining the visitors. This is a straight-forward feature where we simply define the prefecture as the first and second part of the area_name.

One Hot Encoding on the area name and genre name gave us better results compared to Label Encoding. VIF scores were also checked for both area and genre, since the scores were low so no columns were dropped.

Since the "Latitude" and "Longitude" columns have values of the latitude and longitude of the area to which the store belongs, we have added 3 different columns related to them -

The difference of Latitude with the max. Latitude value, the difference of Longitude with the max. Longitude value in the dataset and the sum of Latitude and Longitude of the restaurant. Normalizing the sum of Latitude and Longitude value gave us better results.

These three columns can help the model understand the proximity of the restaurants better and cluster the restaurants in the same area more precisely

V. Training and Results

The following models were attempted for predicting the number of visitors:

i. Linear Regression

ii. K Nearest Neighbours

iii. Light Gradient Boosting Machine

iv. Decision Tree

v. Random Forest

The RMSLE score on the models was the following:

Model Name	RMSLE Score
Linear Regression	0.54005
Decision Tree	0.53515
Light gradient boosting machine	0.52661
Random Forest (trees=500, max depth =8)	0.51690
KNN, k=5	0.46903
KNN, k=4	0.45214
KNN, k=3	0.42977

VI. Conclusion

We would like to conclude that we were able to come up with an efficient model to forecast the number of visitors for different genres of restaurants in different areas of Japan. Such projects have a potential scope to help small businesses and newer restaurants make effective decisions and utilization of their capital and resources for restaurant attendance even for unforeseeable circumstances.

Acknowledgment

We would like to thank Professor G. Srinivas Raghavan, Prof. Neelam Sinha and our Machine Learning Teaching Assistants, Arjun Verma, Tejas Kotha, Shreyas Gupta, for giving us the opportunity to work on the project and help us whenever we were stuck by giving us ideas and resources to learn from. We would also like to thank Team Excelsior for being a great competitor and setting a benchmark time by time for the rest of us which acted as a driving fuel for us to constantly work hard and surpass them. A special thanks to Meghna Dubey and Nihar Kanwar for having a healthy discussion regarding various approaches in pre-processing, feature selection, algorithm optimizations etc. We had a great learning experience while working on the project and the leader board was a great motivation to work on the project.

The competition forced us to read up various articles and papers which gave us ideas and enthusiasm for the project.

REFERENCES

[1] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[2] S. Raschka, "Mlxtend: Providing machine learning and data science utilities and extensions to python's scientific computing stack," The Journal of Open Source Software, vol. 3, no. 24, Apr. 2018. [Online]. Available: http://joss.theoj.org/papers/10.21105/joss.00638

[3] Kaggle Meetup: Recruit Restaurant Visitor Forecasting - https://www.youtube.com/watch?v=6llLC4M3dMo

[4] Junaid Khan, "Recruit Restaurant Visitor Forecasting": https://medium.com/analytics-vidhya/recruit-restaurant-visitor-forecasting-f9ef87ba1073

[5] Sanjay Challal, "Forecasting the visitors for future dates using ML" Available: https://sanjay-c.medium.com/recruit-restaurant-visitor-forecasting-a704cd5432c8